Pattern Recognition Letters 45 (2014) 161–171

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Scalable semi-supervised clustering by spectral kernel learning ☆

M. Soleymani Baghshah [a], F. Afsari [b,*], S. Bagheri Shouraki [c], E. Eslami [d]

[a] Computer Engineering Department, Sharif University of Technology, Tehran, Iran
[b] Computer Engineering Department, Shahid Bahonar University of Kerman, Kerman, Iran
[c] Electrical Engineering Department, Sharif University of Technology, Tehran, Iran
[d] Mathematics and Computer Science, Shahid Bahonar University of Kerman, Kerman, Iran

ARTICLE INFO

Article history:
Received 12 September 2013
Available online 13 April 2014

Keywords:
Kernel learning
Spectral
Scalable
Semi-supervised clustering
Laplacian
Constraint

ABSTRACT

Kernel learning is one of the most important and recent approaches to constrained clustering. Until now many kernel learning methods have been introduced for clustering when side information in the form of pairwise constraints is available. However, almost all of the existing methods either learn a whole kernel matrix or learn a limited number of parameters. Although the non-parametric methods that learn whole kernel matrix can provide capability of finding clusters of arbitrary structures, they are very computationally expensive and these methods are feasible only on small data sets. In this paper, we propose a kernel learning method that shows flexibility in the number of variables between the two extremes of freedom degree. The proposed method uses a spectral embedding to learn a square matrix whose number of rows is the number of dimensions in the embedded space. Therefore, the proposed method shows much higher scalability compared to other methods that learn a kernel matrix. Experimental results on synthetic and real-world data sets show that the performance of the proposed method is generally near to the learning a whole kernel matrix while its time cost is very low compared to these methods.

1. Introduction

Many of the most famous machine learning and pattern recognition methods use a (dis)similarity measure in their methods. Performance of these methods (e.g., SVM, kNN, RBF networks, *k-means*, etc.) is highly dependent on the used measure (i.e., distance metric, kernel function, etc.) and thus it is important to choose a proper measure in them. In the last decade, several methods have been introduced to learn distance or kernel functions (or matrices) instead of predefining them. Although distance and kernel learning methods have been introduced for all categories of supervised [1–9], unsupervised [10–16], and semi-supervised [17–38] tasks, the most interest has been on semi-supervised tasks where limited supervisory information is available.

The existing approaches to distance or kernel learning can be categorized into: (i) explicitly finding a new representation of the data that is more appropriate (e.g., for clustering or classification) and then using an Euclidean distance metric (or inner product similarity measure) in the transformed space. (ii) without explicitly finding transformed data, just finding a proper distance metric or kernel function. On the other hand, existing methods can also be categorized into linear and non-linear methods. Indeed, some of them [1,19–22] learn a linear transformation or equivalently a Mahalanobis distance metric while others try to learn non-linearly transformed data or learn a kernel matrix. Some of the existing non-linear methods [29,33,35–38] indeed learn a linear transformation in a spectrally embedded space (obtained by spectral analysis).

Distance and kernel learning are the most popular approaches for semi-supervised clustering problems. In the semi-supervised clustering tasks, along with unlabeled data, supervisory information in the form of must-links and cannot-link constraints (on some pairs of data) are available. Each must-link constraint implies two data points should be in the same cluster and each cannot-link constraint implies two data points should be in different clusters. So far many Mahalanobis distance learning [19–22] and kernel learning methods [23–35] have been introduced for semi-supervised clustering. Since former methods correspond to learning linear transformations, they are not useful to match arbitrary shapes of clusters. Nonetheless, kernel learning methods have been considered as a more proper approach that can learn non-linearity in the structures of clusters. However, most of the existing kernel

☆ This paper has been recommended for acceptance by Y. Liu.

* Corresponding author. Address: Shahid Bahonar University of Kerman, PO Box: 76169-133, Afzalipoor Square, Kerman, Iran. Tel.: +98 3413202547; fax: +98 341 3235901.

E-mail addresses: soleymani@sharif.edu (M. Soleymani Baghshah), afsari@uk.ac.ir, afsari.f@gmail.com (F. Afsari), sbagheri@sharif.edu (S. Bagheri Shouraki), esfandiar.eslami@uk.ac.ir (E. Eslami).

http://dx.doi.org/10.1016/j.patrec.2014.02.014
0167-8655/

learning methods (for semi-supervised clustering) either show poor flexibility [31–35] or show very low computational efficiency and also poor generalization capability [24,25,28]. In fact, the methods either learn a small number of parameters or learn a whole $n \times n$ matrix where n denotes the number of data and thus the later methods are usually very computationally expensive. Although few recent methods [29,30,38] have tried to learn low rank kernel matrices or learn smaller matrices, their optimization problems are not such suitable for constrained clustering. In this paper, we propose a method that can show flexibility between the two extremes of free parameters. In the proposed method, an $m \times m$ matrix is learned where m can be varied between 1 and n. Moreover, the proposed optimization problem is more suitable for learning a kernel matrix that is used in the kernel *k-means* clustering algorithm [39]. Our method can balance between the available supervisory information and the complexity of the transform to avoid overfitting problems.

In this paper, we propose a novel spectral kernel learning method that uses the data and constraints to find an appropriate kernel matrix. This method is a non-linear metric learning method for semi-supervised clustering that can compromise between flexibility (of finding complex cluster structures) and computational complexity. The proposed method learns an $m \times m$ matrix by solving a Semi-Definite Programming (SDP) problem. The proposed method can also be considered as Mahalanobis metric learning in a representation space obtained according to the geometrical structure of the data. Indeed, we first embed the data in the representation space using spectral embedding and then learn a Mahalanobis metric in this space.The major contributions of the proposed method can be summarized as: first the proposed non-parametric kernel learning method does not need to learn all elements of an $n \times n$ matrix. Indeed, we can specify the freedom degree of kernel learning m (m can be much lower than n). Our method can be scalable to large problems as opposed to the existing non-parametric kernel learning methods [23–28]. Second, although all of the existing kernel learning methods for semi-supervised clustering use the learned kernel in the kernel *k-means* algorithm, they have not attend the relation between the kernel learning phase and the clustering phase accurately. We intend to learn a kernel that matches with the objective function of the *k-means* clustering algorithm. Therefore, our proposed optimization problem in the kernel learning phase is such that it prepares a suitable kernel for the kernel *k-means* clustering algorithm.

The rest of this paper is organized as follows: In Section 2, first we briefly introduce the background on spectral and kernel learning concepts and also the related studies. In Section 3, our proposed method is formulated as an optimization problem and solved accordingly. Experimental results and evaluations are given in Section 4. Finally, Section 5 concludes the paper.

2. Background and related work

In this section, we introduce some preliminaries of spectral and kernel-based methods, followed by some discussion on related work.

2.1. Spectral and kernel-based methods

Let $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ be a set of data points and $G(\mathcal{V}, W)$ be a similarity graph constructed on them. Thus, $\mathcal{V}$ is a set of vertices corresponding to the data points and $\mathbf{W}$ is the weighted adjacency matrix specified according to a similarity measure. There are some popular ways to construct a similarity graph on data points [40]: (i) *ε-nearest-neighbor graph*, in which all the data points whose pairwise distances are smaller than ε are connected; (ii) *k-nearest neighbor graphs*, where v_i is connected to v_j when v_j is among the k-nearest neighbors of v_i; and (iii) *fully connected graph*, where all data points are simply connected with weighted edges. k-nearest neighbors graph is the most famous way of constructing the similarity graph that we also use in this paper.Graph Laplacian matrices that are built based on weight matrix $\mathbf{W}$ have an important role in most of spectral learning methods. The unnormalized and normalized Laplacian matrices of G are respectively defined as $\mathbf{L} = \mathbf{D} - \mathbf{W}$ and $\bar{\mathbf{L}} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}$ where $\mathbf{D}$ is the diagonal matrix whose diagonal elements are the degrees of the vertices of the graph $G\left(D_{ii} = \sum_{j=1}^{n} W_{ij}\right)$. Both $\mathbf{L}$ and $\bar{\mathbf{L}}$ are symmetric and positive semi-definite matrices.

In the last decade, spectral graph theory has been extensively used in clustering [41–43], embedding [44,45], and semi-supervised learning [46–49] problems. The spectral methods are based on properties of matrices like the graph Laplacian. Dhillon et al. [50] showed the theoretical relation between spectral clustering and kernel *k-means*. Moreover, Bengio et al. [51] showed that Laplacian eigenmap can be considered as a KPCA algorithm where kernel matrix is constructed accordingly. Furthermore, many semi-supervised learning methods [46–49] used the Laplacian matrix to find a suitable kernel. Therefore, all of these spectral clustering, embedding, and semi-supervised learning methods can be considered as kernel-based methods where the used kernel is constructed based on the Laplacian matrix or the similarity matrix. According to the spectral graph theory, among the eigenvectors of the Laplacian matrix, the smoother ones are more important (i.e., the eigenspectrum of the graph Laplacian encodes smoothness information over the graph) [40].

A kernel function is a real-valued positive semi-definite function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Indeed, $k(\mathbf{x}, \mathbf{y})$, can be considered as the dot product of the mapped $\mathbf{x}, \mathbf{y}$ in the kernel space (i.e., $k(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$) and a distance metric can be found according to this positive definite kernel. A kernel function is a similarity measure in a space resulted by a feature mapping $\varphi : \mathcal{X} \to \mathcal{F}$ from the input space $\mathcal{X}$ to a feature space $\mathcal{F}$ (where $\mathcal{F}$ is a **Hilbert** space). For the set $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ of data points, the kernel matrix is defined as $\mathbf{K} = \{k(\mathbf{x}_i, \mathbf{x}_j)\}_{1 \leqslant i,\, j \leqslant n}$ where $\mathbf{K}$ is a symmetric positive semi-definite matrix. Indeed, every positive definite and symmetric matrix is a kernel matrix, that is, an inner product matrix in some embedding space and conversely, every kernel matrix is symmetric and positive definite [52].

2.2. Related works

Until now many kernel learning methods have been introduced. Most of them either construct a kernel based on learning small number of coefficients or learn a whole kernel matrix. Some studies [18,31,32] are based on learning only parameters of a parametric kernel like Gaussian kernel function. Some others [3,4,7,9,34–37,53,54] assume kernel matrices to be a linear combination of base matrices constructed beforehand (usually from eigenvectors of a fixed matrix like the graph Laplacian). The most flexible methods [5,23–28] learn an $n \times n$ matrix. However, these methods are usually very computationally expensive and also need large amount of supervisory information to find the large number of free parameters. Thus, the existing methods usually either learn at most n parameters (where n shows the number of data) or learn an $n \times n$ matrix that causes low efficiency and overfitting problems. Few recent methods [29,30] have tried to learn low rank kernel matrices or learn a smaller matrix [38]. However, their optimization problems are not completely suitable for the constrained clustering problem.

Although many kernel (or metric) learning methods have been introduced for semi-supervised clustering, the relation between